entropy

MDPI

Article

The Kullback–Leibler Information Function for Infinite Measures

Victor Bakhtin [1,*] and Edvard Sokal [2]

[1] Department of Mathematics, IT and Landscape Architecture, John Paul II Catholic University of Lublin, Konstantynuv Str. 1H, 20-708 Lublin, Poland

[2] Department of Mechanics and Mathematics, Belarusian State University, Nezavisimosti Ave. 4, 220030 Minsk, Belarus; edward.e.sokol@gmail.com

* Correspondence: bakhtin@tut.by or bakhtin@kul.lublin.pl; Tel.: +375-25-934-3780

Academic Editor: Raúl Alcaraz Martínez
Received: 21 July 2016; Accepted: 12 December 2016; Published: 15 December 2016

Abstract: In this paper, we introduce the Kullback–Leibler information function $\rho(\nu, \mu)$ and prove the local large deviation principle for σ-finite measures μ and finitely additive probability measures ν. In particular, the entropy of a continuous probability distribution ν on the real axis is interpreted as the exponential rate of asymptotics for the Lebesgue measure of the set of those samples that generate empirical measures close to ν in a suitable fine topology.

Keywords: Kullback–Leibler information function; entropy; large deviation principle; empirical measure; fine topology; spectral potential

MSC: 28D20; 60F10

1. Introduction

Let P be a continuous probability distribution on the real axis with density $\varphi(x) = dP(x)/dx$. Its *entropy* is defined as

$$H(P) = -\int_{\mathbb{R}} \varphi(x) \ln \varphi(x)\, dx. \tag{1}$$

What is the substantive sense of $H(P)$? More precisely, does there exist a mathematical object whose natural quantitative magnitude (e.g., volume) is a certain function of the entropy?

Traditionally, entropy is treated as a measure of disorder. However, this explanation does not answer the question stated above because it does not establish a relationship between entropy and any other quantitative characteristic of disorder that can be defined and measured regardless of the entropy.

To illustrate the problem, consider the entropy of a discrete distribution $P = (p_1, \dots, p_r)$,

$$H(P) = -\sum_i p_i \ln p_i. \tag{2}$$

Its substantive meaning is well known. Namely, let $X = \{1, \dots, r\}$ be a finite alphabet. Then, the set of those words $(x_1, \dots, x_n) \in X^n$ of length $n \gg 1$ in which every letter $i \in X$ occurs with mean frequency close to p_i has cardinality of order $e^{nH(P)}$ (this follows from the Shannon–McMillan–Breiman theorem (see [1,2])). Thus, the entropy of a discrete distribution determines the exponential rate for the number of those words of length n in which letters occur with prescribed frequencies.

Can we say anything of that sort about the entropy of a continuous distribution? It turns out—yes. Indeed, from Theorem 3 stated below, it follows that entropy (1) determines the exponential rate for the *Lebesgue measure* of the set of sequences $(x_1, \dots, x_n) \in \mathbb{R}^n$ of length $n \gg 1$ that generate empirical measures on $\mathbb{R}$ close to P. The proximity of distributions should be understood here in the sense of a

fine topology, which is defined in the same way as the weak topology, but with the use of integrable functions instead of bounded ones.

For example, if P is the exponential distribution with density $\varphi(x) = \lambda e^{-\lambda x}$, $x \geq 0$, then

$$H(P) = -\int_0^{+\infty} \lambda e^{-\lambda x}(\ln \lambda - \lambda x)\, dx = 1 - \ln \lambda,$$

and so the set of sequences $(x_1, \ldots, x_n) \in \mathbb{R}^n$ of length $n \gg 1$ that generate empirical measures close to P (in the fine topology) has Lebesgue measure of order $e^{nH(P)} = (e/\lambda)^n$.

Another example: for the Gaussian distribution P with density

$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-a)^2/2\sigma^2},$$

we get

$$H(P) = -\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-a)^2/2\sigma^2} \left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-a)^2}{2\sigma^2} \right) dx = \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2},$$

and the set of sequences $(x_1, \ldots, x_n) \in \mathbb{R}^n$ of length $n \gg 1$ that generate empirical measures close to P (in the fine topology) has Lebesgue measure of order $e^{nH(P)} = (2\pi\sigma^2 e)^{n/2}$.

These examples are based on the presentation of entropy (1) in the form $H(P) = -\rho(P, Q)$, where Q is the Lebesgue measure on the real axis and $\rho(P, Q)$ is the Kullback–Leibler information function:

$$\rho(P, Q) = \int_{\mathbb{R}} \varphi(x) \ln \varphi(x)\, dQ(x), \qquad \varphi(x) = \frac{dP(x)}{dQ(x)}, \tag{3}$$

as well as on a certain generalization of the so-called local large deviation principle.

Let P and Q be two probability distributions on a space X. Roughly speaking, the local large deviation principle asserts that the measure Q^n of the set of sequences $(x_1, \ldots, x_n) \in X^n$ that generate empirical measures close to P has exponential order $e^{-n\rho(P,Q)}$, provided $n \to +\infty$.

As far as we know this principle was first proven by Sanov for a pair of continuous probability distributions on the real axis in [3]. Later, it was extended to the general metric spaces (see, for example, [4–7]), abstract measurable spaces (see [8–10]), and spaces of trajectories of various stochastic processes (see [11–19]).

It should be mentioned that different authors called the function $\rho(P, Q)$ in different ways: the Kullback–Leibler information function [4], the relative entropy [6], the rate function [5,7,15], the Kullback–Leibler divergence, the action functional [16], and the Kullback–Leibler distance [20] (though, of course, it is nonsymmetric and hence not a metric at all). For brevity, in the sequel, we will prefer the term "Kullback action" rather than any of the listed above.

Until recently, the Kullback action and the local large deviation principle were studied only in the case when both arguments P, Q were probability distributions. Only recently, in papers [9,10], was the measure Q allowed to be no more than finite and positive, and the measure P was allowed to be finitely additive, and, moreover, real-valued. Unfortunately, this is still insufficient for the interpretation of entropy (1) because the Lebesgue measure on the real axis is *infinite*. Therefore, it is highly desirable to define properly the Kullback action and to obtain a generalization of the local large deviation principle for infinite measures Q. Our main result is the solution of this problem.

It turns out that at least two different ways of generalization are possible. The first approach is based on the use of the fine topology in the space of probability distributions. This is presented in Theorem 3. In the second approach, the whole space X is replaced by its certain part Y of finite measure Q, and the distribution P is replaced by its conditional distribution P_Y on Y. Thereby, the problem reduces to the case of finite measures. This approach is implemented in Theorems 4 and 5.

In fact, it makes sense to consider finitely additive probability distributions P as well since some sequences of empirical measures may converge to finitely additive distributions. In such a case, the Kullback action can take values $+\infty$ or $-\infty$ only (Theorem 6). The corresponding versions of the large deviation principle for finitely additive measures P are presented in Theorems 7 and 8.

First results on the large deviation principle for infinite measures were obtained in [21,22], where a countable set X and the "counting" measure Q (such that $Q(x) = 1$ for all $x \in X$) were considered. In such a case, the Kullback action $\rho(P, Q)$ coincides (up to the sign) with entropy (2). It was revealed in [21,22] that, for the "counting" measure Q on the countable space X, the ordinary form of the large deviation principle, formulated in terms of the weak topology, fails and so one should use the fine topology instead.

The paper is organized as follows. In the next section we recall the local large deviation principle for finite measures (Theorem 1). In Section 3, we define the Kullback action $\rho(\nu, \mu)$ as the Legendre dual functional to the so-called spectral potential $\lambda(\varphi, \mu)$ and formulate two variants of the large deviation principle for the case of σ-finite measure μ (Theorems 3–5). These theorems are proven in Sections 4–7. In Section 8, we formulate two variants of the large deviation principle for σ-finite measures μ and finitely additive probability distributions ν (Theorems 7 and 8). Theorem 6 states that, in fact, $\rho(\nu, \mu)$ turns into $+\infty$ or $-\infty$ if the measure ν has no density with respect to μ. It is proven in Section 9. The final Section 10 contains proofs of Theorems 7 and 8.

2. The Kullback Action for Finite Measures

Let us consider an arbitrary set X supplied with a σ-field $\mathfrak{A}$ of its subsets. In what follows by "measures" we mean only *nonnegative* measures on the measurable space $(X, \mathfrak{A})$.

We will use the following notation:

$B(X)$ — all bounded measurable functions $f\colon (X, \mathfrak{A}) \to \mathbb{R}$;

$M(X)$ — all finite measures on $(X, \mathfrak{A})$;

$M_1(X)$ — all probability measures (distributions) on $(X, \mathfrak{A})$;

$M_\sigma(X)$ — all σ-finite measures on $(X, \mathfrak{A})$.

Evidently,

$$M_1(X) \subset M(X) \subset M_\sigma(X).$$

Suppose that $\nu, \mu \in M_\sigma(X)$ and the measure ν is absolutely continuous with respect to μ. Then, by the Radon–Nikodym theorem, ν can be presented in the form $\nu = \varphi\mu$, where φ is a nonnegative measurable function, which is called the density of ν with respect to μ and denoted as $\varphi = d\nu/d\mu$. This function is uniquely defined up to a set of zero measure μ.

The *Kullback action* $\rho(\nu, \mu)$ is a function of a probability measure $\nu \in M_1(X)$ and a finite measure $\mu \in M(X)$ defined in the following way: if ν is absolutely continuous with respect to μ, then

$$\rho(\nu, \mu) = \int_X \varphi \ln \varphi \, d\mu, \qquad \varphi = \frac{d\nu}{d\mu}, \tag{4}$$

and $\rho(\nu, \mu) = +\infty$, otherwise. In (4), we set $\varphi \ln \varphi = 0$ for $\varphi = 0$. Therefore, $\rho(\nu, \mu)$ belongs to the interval $(-\infty, +\infty]$.

With each finite sequence $x = (x_1, \dots, x_n) \in X^n$, we associate an *empirical measure* $\delta_{x,n} \in M_1(X)$ that is supported on the set $\{x_1, \dots, x_n\}$ and assigns to each x_i the measure $1/n$. The expectation of any function $f\colon X \to \mathbb{R}$ with respect to this empirical measure looks like

$$\delta_{x,n}[f] = \frac{f(x_1) + \dots + f(x_n)}{n}.$$

Let us fix any probability measure $\mu \in M_1(X)$. If the points $x_i \in X$ are treated as independent random variables with common distribution μ, then the empirical measure $\delta_{x,n}$ becomes a random

variable itself, taking values in $M_1(X)$. We will be interested in the asymptotics of its distribution. It turns out that, at a first approximation, this asymptotics is exponential with the exponent $-n\rho(\nu,\mu)$.

To describe the asymptotics of the empirical measures distribution, we need two topologies on the space $M_1(X)$. The first one is the weak topology generated by neighborhoods of the form

$$O(\mu) = \left\{\nu \in M_1(X) : \left|\int_X f_i\,d\nu - \int_X f_i\,d\mu\right| < \varepsilon,\ \ i = 1,\dots,k\right\}, \quad \mu \in M_1(X), \tag{5}$$

where $f_1,\dots,f_k \in B(X)$ and $\varepsilon > 0$. The second topology is generated by neighborhoods of the same form (5) but with functions $f_1,\dots,f_k \in L^1(X,\mu)$ therein. In addition, it is supposed in this case that $O(\mu)$ contains only those measures ν for which all integrals $\int_X f_i\,d\nu$ do exist. This topology will be referred to as the *fine topology*. It is useful because it enables us to formulate the usual law of large numbers in the next form: for any probability distribution $\mu \in M_1(X)$, the sequence of empirical measures $\delta_{x,n}$ converges to μ in the probability in the fine topology. On the other hand, a shortcoming of the fine topology is the fact that, with respect to it, the affine map $t \mapsto (1-t)\mu_0 + t\mu_1$, where $t \in [0,1]$, may be discontinuous at the ends of the segment $[0,1]$.

It is easy to see that the fine topology on $M_1(X)$ contains the weak one, but the converse, in general, does not take place.

For any nonnegative measure μ on X, denote by μ^n its Cartesian power supported on X^n. The next theorem describes asymptotics of the empirical measures distribution.

Theorem 1 (the local large deviation principle for finite measures)**.** *For any measures $\nu \in M_1(X)$, $\mu \in M(X)$, and number $\varepsilon > 0$, there exists a weak neighborhood $O(\nu) \subset M_1(X)$ such that*

$$\mu^n\{x = (x_1,\dots,x_n) \in X^n \mid \delta_{x,n} \in O(\nu)\} \le e^{-n(\rho(\nu,\mu)-\varepsilon)}, \qquad n = 1,2,3,\dots \tag{6}$$

On the other hand, for any measures $\nu \in M_1(X)$, $\mu \in M(X)$, number $\varepsilon > 0$, and any fine neighborhood $O(\nu) \subset M_1(X)$, the following estimate holds for all large enough n:

$$\mu^n\{x = (x_1,\dots,x_n) \in X^n \mid \delta_{x,n} \in O(\nu)\} \ge e^{-n(\rho(\nu,\mu)+\varepsilon)}. \tag{7}$$

In the case of a metric space X supplied with a Borel σ-field, the neighborhood $O(\nu)$ in (6) can be chosen from the weak topology generated by bounded continuous functions.

Remark 1. *When $\rho(\nu,\mu) = +\infty$, the difference $\rho(\nu,\mu) - \varepsilon$ in (6) should be replaced by $1/\varepsilon$.*

Remark 2. *So long as each weak neighborhood in $M_1(X)$ belongs to the fine topology, estimates (6) and (7) complement each other: the coefficient $\rho(\nu,\mu)$ cannot be increased in (6) and cannot be decreased in (7).*

Remark 3. *Theorem 1 is also true for* finitely additive *probability distributions ν on the space X if we set $\rho(\nu,\mu) = +\infty$ in such a case (see [9]).*

It is worth mentioning that, until recently, the absolute majority of papers on the large deviation principle dealt with random variables in Polish space (i.e., complete separable metric space), and only a few of them treated random variables in a topological space (see, for example, [4]), or in a measurable space in which the σ-field is generated by open balls and does not necessarily contain Borel sets (see [7], Section 7). In addition, only countably additive probability distributions ν and μ were considered as arguments of the Kullback action. Theorem 1 for an arbitrary measurable space X, finitely additive measures ν and nonnormalized measures μ was first proven in [9], and its generalization for finitely additive measures μ was proven in [10].

3. The Kullback Action for σ-Finite Measures

The shortcoming of Theorem 1 is that it does not involve the case of infinite measure μ. In particular, it does not explain any sense of entropy (1) of an absolutely continuous probability

distribution on the real axis. Unfortunately, the direct extension of Theorem 1 on infinite measures μ is wrong. The next example demonstrates this.

Example ([22]). *Let X be a countable set supplied with the discrete σ-field and μ be the counting measure on X (such that $\mu(x) = 1$ for every $x \in X$). Consider a topology on the space of probability distributions $M_1(X)$ generated by the neighborhoods*

$$O_\varepsilon(\nu) = \Big\{\delta \in M_1(X) : \sum_{x\in X} |\delta(x) - \nu(x)| < \varepsilon\Big\}, \qquad \nu \in M_1(X),\ \ \varepsilon > 0 \tag{8}$$

(in other words, the topology of $L^1(X,\mu)$). Then, for any neighborhood (8) and any number $C > 0$, there exists a finite subset $X_0 \subset X$ such that, for all n large enough,

$$\mu^n\big\{x = (x_1,\dots,x_n) \in X_0^n \bigm| \delta_{x,n} \in O(\nu)\big\} > e^{Cn}. \tag{9}$$

The topology on $M_1(X)$ under consideration contains the weak topology generated by functions from $B(X)$. It follows that, for $C > -\rho(\nu,\mu)$, estimate (9) contradicts (6), and hence the latter cannot take place.

It turns out that, to extend Theorem 1 on σ-finite measures μ, it is enough to replace the weak neighborhood in (6) with a fine one. This is the main result of the paper. Its exact formulation is given in Theorem 3 below.

We also propose one more approach to extend Theorem 1, using only weak topology. Its idea is to replace the space X in estimates (6) and (7) by a large enough subset $Y \subset X$ of finite measure $\mu(Y)$, and to replace the probability measure $\nu \in M_1(X)$ by its conditional distribution on Y. The corresponding results are stated in Theorems 4 and 5 below.

In order to describe asymptotics of the empirical measures distribution correctly in the case of σ-finite measure μ, the definition of the Kullback action should be modified. To this end, we have to introduce the notion of a spectral potential.

Denote by $\bar{B}(X)$ the set of all bounded above measurable functions on a measurable space $(X,\mathfrak{A})$. The *spectral potential* is the nonlinear functional

$$\lambda(\varphi,\mu) = \ln\int_X e^{\varphi}\,d\mu, \quad \varphi \in \bar{B}(X),\ \ \mu \in M_\sigma(X).$$

If the integral in this formula diverges, then we set $\lambda(\varphi,\mu) = +\infty$. Thus, $\lambda(\varphi,\mu)$ can take values in the interval $(-\infty,+\infty]$.

For brevity, let us introduce the notation

$$\nu[f] = \int_X f\,d\nu,$$

where $\nu \in M_1(x)$ and $f \in \bar{B}(X)$. If the integral diverges, then we put $\nu[f] = -\infty$.

Now, we define the *Kullback action* $\rho(\nu,\mu)$ as a function of the pair of arguments $\nu \in M_1(X)$ and $\mu \in M_\sigma(X)$ as follows:

$$\rho(\nu,\mu) = \begin{cases} +\infty, & \text{if } \exists\, A \in \mathfrak{A}\colon\ \mu(A) = 0,\ \nu(A) > 0, \\ \sup\limits_{\psi\in\bar{B}(X)} \big\{\nu[\psi] - \lambda(\psi,\mu)\big\}, & \text{if } \nexists\, A \in \mathfrak{A}\colon\ \mu(A) = 0,\ \nu(A) > 0. \end{cases} \tag{10}$$

The next theorem shows, in particular, that in the case of a finite measure μ this definition coincides with the previous one (4).

Theorem 2. *If a probability distribution $\nu \in M_1(X)$ is absolutely continuous with respect to $\mu \in M_\sigma(X)$ and $d\nu/d\mu = \varphi$, then*

$$\rho(\nu,\mu) = \int_X \varphi \ln \varphi \, d\mu, \qquad \text{if} \quad \int_{\varphi<1} \varphi \ln \varphi \, d\mu > -\infty, \tag{11}$$

$$\rho(\nu,\mu) = -\infty, \qquad \text{if} \quad \int_{\varphi<1} \varphi \ln \varphi \, d\mu = -\infty. \tag{12}$$

In particular, for the finite measure μ, the alternative (11) *takes place.*

The following theorem is our main result for the case of countably additive distributions.

Theorem 3 (the local large deviation principle for infinite measures)**.** *For any measures $\nu \in M_1(X)$, $\mu \in M_\sigma(X)$, and number $\varepsilon > 0$, there exists a fine neighborhood $O(\nu) \subset M_1(X)$ such that*

$$\mu^n\{x = (x_1, \dots, x_n) \in X^n \mid \delta_{x,n} \in O(\nu)\} \le e^{-n(\rho(\nu,\mu)-\varepsilon)}, \qquad n = 1, 2, 3, \dots \tag{13}$$

On the other hand, for any measures $\nu \in M_1(X)$, $\mu \in M_\sigma(X)$, number $\varepsilon > 0$, and any fine neighborhood $O(\nu) \subset M_1(X)$, the following estimate holds for all large enough n:

$$\mu^n\{x = (x_1, \dots, x_n) \in X^n \mid \delta_{x,n} \in O(\nu)\} \ge e^{-n(\rho(\nu,\mu)+\varepsilon)}. \tag{14}$$

If $\rho(\nu,\mu) = +\infty$, then the difference $\rho(\nu,\mu) - \varepsilon$ in (13) *should be replaced by $1/\varepsilon$, and if $\rho(\nu,\mu) = -\infty$ then the sum $\rho(\nu,\mu) + \varepsilon$ in* (14) *should be replaced by $-1/\varepsilon$.*

Let us also formulate the local large deviation principle in terms of weak neighborhoods.

For any probability measure $\nu \in M_1(X)$ and any measurable subset $Y \subset X$ with $\nu(Y) > 0$, define a conditional measure $\nu_Y \in M_1(X)$ according to the formula

$$\nu_Y(A) = \frac{\nu(A \cap Y)}{\nu(Y)}, \qquad A \in \mathfrak{A}.$$

It is easily seen that the measure ν can be approximated by the conditional measures ν_Y, where $\mu(Y) < +\infty$, in the fine topology (and all the more in the weak one). Therefore, it can make sense to replace fine neighborhoods of ν in Theorem 3 by weak neighborhoods of close conditional measures ν_Y.

We will say that the Kullback action $\rho(\nu,\mu)$ is *well-defined* if ν has a density $\varphi = d\nu/d\mu$, and, in addition, at least one of the two integrals

$$\int_{\varphi<1} \varphi \ln \varphi \, d\mu, \qquad \int_{\varphi\ge1} \varphi \ln \varphi \, d\mu \tag{15}$$

is finite. In all other cases (i.e., when both integrals (15) are infinite or the measure ν has no density with respect to μ), we will say that the Kullback action is *ill-defined*.

Theorem 4. *Suppose that, for some measures $\nu \in M_1(X)$ and $\mu \in M_\sigma(X)$, the Kullback action $\rho(\nu,\mu)$ is well-defined. Then, for any number $\varepsilon > 0$, there exists a set $X_\varepsilon \in \mathfrak{A}$ with $\mu(X_\varepsilon) < +\infty$ such that for any $Y \in \mathfrak{A}$ containing X_ε and having a finite measure $\mu(Y)$:*

(*a*) *there exists a weak neighborhood $O(\nu_Y) \subset M_1(Y)$ satisfying the estimate*

$$\mu^n\{x = (x_1, \dots, x_n) \in Y^n \mid \delta_{x,n} \in O(\nu_Y)\} \le e^{-n(\rho(\nu,\mu)-\varepsilon)}, \qquad n = 1, 2, 3, \dots; \tag{16}$$

(*b*) *for any fine neighborhood* $O(\nu_Y) \subset M_1(Y)$ *and all large enough* n,

$$\mu^n\{x = (x_1, \dots, x_n) \in Y^n \mid \delta_{x,n} \in O(\nu_Y)\} \geq e^{-n(\rho(\nu,\mu)+\varepsilon)}. \tag{17}$$

In addition, for any $\varepsilon > 0$ *and any fine neighborhood* $O(\nu) \subset M_1(X)$, *there exists a set* $Y \in \mathfrak{A}$ *with* $\mu(Y) < +\infty$ *such that for all large enough* n,

$$\mu^n\{x = (x_1, \dots, x_n) \in Y^n \mid \delta_{x,n} \in O(\nu)\} \geq e^{-n(\rho(\nu,\mu)+\varepsilon)}. \tag{18}$$

Theorem 5. *Suppose that for some measures* $\nu \in M_1(X)$ *and* $\mu \in M_\sigma(X)$, *the Kullback action* $\rho(\nu,\mu)$ *is ill-defined. Then, there exists a set* $X_0 \in \mathfrak{A}$ *with* $\mu(X_0) < +\infty$, *such that, for any* $Y \in \mathfrak{A}$ *containing* X_0 *and having a finite measure* $\mu(Y)$, *and any* $\varepsilon > 0$, *there exists a weak neighborhood* $O(\nu_Y) \subset M_1(Y)$ *satisfying the estimate*

$$\mu^n\{x = (x_1, \dots, x_n) \in Y^n \mid \delta_{x,n} \in O(\nu_Y)\} \leq e^{-n/\varepsilon}, \qquad n = 1, 2, 3, \dots \tag{19}$$

It is worth mentioning that, under conditions of Theorem 5, the equality $\rho(\nu,\mu) = -\infty$ may take place. In such a case, estimates (19) and (14) have opposite senses. Nevertheless, there is no contradiction here because the sets in these estimates are different.

4. Proof of Theorem 2

Recall that, under conditions of Theorem 2, the measure $\nu \in M_1(X)$ is absolutely continuous with respect to $\mu \in M_\sigma(X)$ and has a density $\varphi = d\nu/d\mu$. First of all, we will prove that for any function $\psi \in \bar{B}(X)$,

$$\nu[\psi] - \lambda(\psi,\mu) \leq \begin{cases} \displaystyle\int_X \varphi \ln \varphi \, d\mu, & \text{if } \displaystyle\int_{\varphi<1} \varphi \ln \varphi \, d\mu > -\infty, \\ -\infty, & \text{if } \displaystyle\int_{\varphi<1} \varphi \ln \varphi \, d\mu = -\infty. \end{cases} \tag{20}$$

If at least one of the expressions $\nu[\psi]$ or $\lambda(\psi,\mu)$ takes the infinite value allowed to it, then the left-hand side of (20) turns into $-\infty$, and so the inequality is true. Thus, it is enough to consider the case of finite $\nu[\psi]$ and $\lambda(\psi,\mu)$.

Suppose first that

$$\int_{\varphi<1} \varphi \ln \varphi \, d\mu > -\infty.$$

For any $\varepsilon > 0$, define the set

$$A_\varepsilon = \{x \in X : \varepsilon < \varphi(x) < 1/\varepsilon, \ \psi(x) > -1/\varepsilon\}$$

and the conditional distribution ν_ε on it:

$$\nu_\varepsilon(B) = \frac{\nu(B \cap A_\varepsilon)}{\nu(A_\varepsilon)}, \qquad B \in \mathfrak{A}. \tag{21}$$

Evidently, ν_ε has the density

$$\varphi_\varepsilon = \frac{d\nu_\varepsilon}{d\mu} = \frac{\chi_\varepsilon \varphi}{\nu(A_\varepsilon)}, \tag{22}$$

where χ_ε is the characteristic function of A_ε.

From elementary properties of integrals, it follows that

$$\lambda(\psi,\mu) = \ln \int_X e^\psi \, d\mu \geq \ln \int_{A_\varepsilon} e^\psi \, d\mu = \ln \int_{A_\varepsilon} e^{\psi - \ln \varphi_\varepsilon} \, d\nu_\varepsilon \tag{23}$$

$$\geq \int_{A_\varepsilon} (\psi - \ln \varphi_\varepsilon)\, d\nu_\varepsilon = \int_{A_\varepsilon} (\psi - \ln \varphi + \ln \nu(A_\varepsilon))\, \frac{d\nu}{\nu(A_\varepsilon)} \tag{24}$$

$$= \frac{1}{\nu(A_\varepsilon)} \int_{A_\varepsilon} \psi\, d\nu - \frac{1}{\nu(A_\varepsilon)} \int_{A_\varepsilon} \varphi \ln \varphi\, d\mu + \ln \nu(A_\varepsilon) \tag{25}$$

(in the passage from (23) to (24), Jensen's inequality is used). If $\varepsilon \to 0$, the expression in (25) converges to

$$\nu[\psi] - \int_X \varphi \ln \varphi\, d\mu.$$

Therefore, (23)–(25) imply the first case of (20) in the limit.

Now, suppose that $\nu[\psi]$ and $\lambda(\psi, \mu)$ are finite and

$$\int_{\varphi<1} \varphi \ln \varphi\, d\mu = -\infty. \tag{26}$$

Consider the sets

$$A_\varepsilon = \{x \in X : \varepsilon < \varphi(x) < 1\}, \qquad \varepsilon \geq 0.$$

As before, define the conditional distributions ν_ε and densities φ_ε by means of (21) and (22). Then, calculations (23)–(25) still hold, but the expression in (25) converges now to the limit

$$\frac{1}{\nu(A_0)} \int_{A_0} \psi\, d\nu - \frac{1}{\nu(A_0)} \int_{A_0} \varphi \ln \varphi\, d\mu + \ln \nu(A_0). \tag{27}$$

In the situation under consideration, the first and the third summands in (27) are finite, while the second one turns into $+\infty$. Therefore, from (23)–(25), it follows that $\lambda(\psi, \mu) = +\infty$, which contradicts the assumption about finiteness of $\lambda(\psi, \mu)$. Thus, in the situation when both $\nu[\psi]$ and $\lambda(\psi, \mu)$ are finite, equality (26) cannot take place. Thereby, inequality (20) is completely proven.

To finish the proof of Theorem 2, it is enough to verify the equality

$$\sup_{\psi \in \bar{B}(X)} \{\nu[\psi] - \lambda(\psi, \mu)\} = \begin{cases} \int_X \varphi \ln \varphi\, d\mu, & \text{if } \int_{\varphi<1} \varphi \ln \varphi\, d\mu > -\infty, \\ -\infty, & \text{if } \int_{\varphi<1} \varphi \ln \varphi\, d\mu = -\infty. \end{cases} \tag{28}$$

By virtue of (20) the left-hand side of (28) does not exceed the right-hand one. If the right-hand side of (28) equals $-\infty$, then the equality is trivial. Consider the case when the right-hand side of (28) is greater than $-\infty$. By σ-finiteness of μ, there exists a function $\eta \in \bar{B}(X)$ such that the integral $\int_X e^\eta\, d\mu$ is also finite. Consider the family of functions

$$\psi_t(x) = \begin{cases} \eta(x) - t, & \text{if } \varphi(x) = 0, \\ \ln \varphi(x), & \text{if } 0 < \varphi(x) \leq e^t, \\ t, & \text{if } \varphi(x) > e^t, \end{cases} \qquad t \in \mathbb{R},$$

Obviously, $\psi_t \in \bar{B}(X)$, and if t goes to $+\infty$, then

$$\int_X e^{\psi_t}\, d\mu = \int_{\varphi=0} e^{\eta - t}\, d\mu + \int_{0<\varphi\leq e^t} \varphi\, d\mu + \int_{\varphi>e^t} e^t\, d\mu \longrightarrow \int_X \varphi\, d\mu = 1,$$

$$\nu[\psi_t] = \int_{0<\varphi\leq e^t} \varphi \ln \varphi\, d\mu + \int_{\varphi>e^t} t\varphi\, d\mu \longrightarrow \int_X \varphi \ln \varphi\, d\mu,$$

$$\nu[\psi_t] - \lambda(\psi_t, \mu) = \nu[\psi_t] - \ln \int_X e^{\psi_t}\, d\mu \longrightarrow \int_X \varphi \ln \varphi\, d\mu.$$

It follows that the supremum in the left-hand side of (28) coincides with the right-hand side. □

5. Proof of the First Part of Theorem 3

At first, suppose that there exists a measurable set A with $\mu(A) = 0$ and $\nu(A) > 0$. Then, by definition $\rho(\nu, \mu) = +\infty$. Denote by χ_A the characteristic function of A. Define a fine neighborhood (in fact a weak one) of the measure ν as follows:

$$O(\nu) = \{\delta \in M_1(X) : \delta[\chi_A] > \nu[\chi_A]/2\}, \quad \text{where } \nu[\chi_A] = \nu(A) > 0.$$

If a sequence $x = (x_1, \dots, x_n) \in X^n$ satisfies the condition $\delta_{x,n} \in O(\nu)$, then $\delta_{x,n}[\chi_A] > 0$. This implies that at least one of the points x_i belongs to A. Therefore,

$$\mu^n\{x = (x_1, \dots, x_n) \in X^n \mid \delta_{x,n} \in O(\nu)\} = 0,$$

which implies (13), where $\rho(\nu, \mu) - \varepsilon$ is replaced by $1/\varepsilon$ by convention.

Now suppose that the second case of formula (10) holds true:

$$\rho(\nu, \mu) = \sup_{\psi \in \bar{B}(X)} \{\nu[\psi] - \lambda(\psi, \mu)\}. \tag{29}$$

If $\rho(\nu, \mu) = -\infty$, then estimate (13) is trivial. Thus, let $\rho(\nu, \mu) > -\infty$. In this case, (29) implies that, for any $\varepsilon > 0$, there exists a function $\psi \in \bar{B}(X)$ that satisfies

$$\rho(\nu, \mu) - \varepsilon/2 < \nu[\psi] - \lambda(\psi, \mu). \tag{30}$$

From this inequality, it follows automatically that $\nu[\psi] > -\infty$ and $\lambda(\psi, \mu) < +\infty$.

Consider the probability distribution $\mu_\psi = e^{\psi - \lambda(\psi,\mu)}\mu$. For any sequence $x = (x_1, \dots, x_n) \in X^n$, we have

$$\frac{d\mu^n(x)}{d\mu_\psi^n(x)} = \prod_{i=1}^n \frac{d\mu(x_i)}{d\mu_\psi(x_i)} = \prod_{i=1}^n e^{\lambda(\psi,\mu) - \psi(x_i)} = e^{n(\lambda(\psi,\mu) - \delta_{x,n}[\psi])}. \tag{31}$$

Define a fine neighborhood of the measure $\nu \in M_1(X)$ as follows:

$$O(\nu) = \{\delta \in M_1(X) : \delta[\psi] > \nu[\psi] - \varepsilon/2\}.$$

Then, under the condition $\delta_{x,n} \in O(\nu)$, it follows from (30) and (31) that

$$\frac{d\mu^n(x)}{d\mu_\psi^n(x)} = e^{n(\lambda(\psi,\mu) - \delta_{x,n}[\psi])} < e^{n(\lambda(\psi,\mu) - \nu[\psi] + \varepsilon/2)} < e^{n(-\rho(\nu,\mu) + \varepsilon)}.$$

Next, since the measure μ_ψ^n is probabilistic,

$$\mu^n\{x \in X^n \mid \delta_{x,n} \in O(\nu)\} = \int\limits_{\delta_{x,n} \in O(\nu)} d\mu^n(x) \le \int\limits_{\delta_{x,n} \in O(\nu)} e^{n(-\rho(\nu,\mu)+\varepsilon)}\, d\mu_\psi^n(x) \le e^{n(-\rho(\nu,\mu)+\varepsilon)}.$$

Thus, inequality (13) is proven in all cases. □

6. Proof of the Second Part of Theorem 3

Now let us proceed to estimate (14). It is trivial if $\rho(\nu, \mu) = +\infty$. Thus, in the sequel, we may suppose that $\rho(\nu, \mu) \in [-\infty, +\infty)$. Then, (10) implies that ν is absolutely continuous with respect to μ and has a density $\varphi = d\nu/d\mu$.

First, consider the case of finite $\rho(\nu,\mu)$. Then, Theorem 2 implies

$$\rho(\nu,\mu)=\int_X \varphi\ln\varphi\,d\mu=\int_X \ln\varphi\,d\nu=\nu[\ln\varphi].$$

Fix any $\varepsilon>0$ and any fine neighborhood $O(\nu)\subset M_1(X)$. Consider the sets

$$Y_n=\left\{x\in X^n \;\middle|\; \delta_{x,n}\in O(\nu),\ \ \left|\delta_{x,n}[\ln\varphi]-\nu[\ln\varphi]\right|<\varepsilon/2\right\}$$

(in the latter inequality, it is supposed that each element of the sequence $x=(x_1,\dots,x_n)$ satisfies the condition $\varphi(x_i)>0$). Note that, for $x\in Y_n$,

$$\frac{d\mu^n(x)}{d\nu^n(x)}=\prod_{i=1}^n\frac{d\mu(x_i)}{d\nu(x_i)}=\prod_{i=1}^n\frac{1}{\varphi(x_i)}=e^{-n\delta_{x,n}[\ln\varphi]}>e^{-n(\nu[\ln\varphi]+\varepsilon/2)}.$$

Hence,

$$\mu^n(Y_n)=\int_{Y_n}d\mu^n(x)\ge\int_{Y_n}e^{-n(\nu[\ln\varphi]+\varepsilon/2)}\,d\nu^n(x)=e^{-n(\rho(\nu,\mu)+\varepsilon/2)}\nu^n(Y_n). \tag{32}$$

By the law of large numbers, $\nu^n(Y_n)\to 1$. Thus, (32) implies (14).

Now, suppose that $\rho(\nu,\mu)=-\infty$. Then, by Theorem 2,

$$\int_{\varphi<1}\varphi\ln\varphi\,d\mu=-\infty. \tag{33}$$

Divide the whole space X into two parts: $X=X^-\sqcup X^+$, where

$$X^-=\{x\in X\mid\varphi(x)<1\},\qquad X^+=\{x\in X\mid\varphi(x)\ge 1\}.$$

Set $X_k^+=\{x\in X^+\mid 1\le\varphi(x)\le k\}$. Evidently, $X^+=\bigcup_k X_k^+$ and

$$\mu(X_k^+)\le\mu(X^+)=\int_{X^+}d\mu=\int_{X^+}\frac{d\nu}{\varphi}\le\int_{X^+}d\nu=\nu(X^+)\le 1. \tag{34}$$

Then, construct a sequence of embedded sets $X_1^-\subset X_2^-\subset X_3^-\subset\cdots$ with $\mu(X_k^-)<+\infty$, such that $X^-=\bigcup_k X_k^-$, and, at the same time,

$$\int_{Y_k}\varphi\ln\varphi\,d\mu\to-\infty,\quad\text{where }\ Y_k=X_k^-\cup X_k^+. \tag{35}$$

Such construction is possible due to (33) and (34). Evidently, $Y_1\subset Y_2\subset Y_3\subset\cdots$, and each Y_k is of finite measure μ, and their union gives the whole X.

Denote by ν_k the conditional distribution of ν on Y_k:

$$\nu_k(A)=\frac{\nu(A\cap Y_k)}{\nu(Y_k)},\qquad A\in\mathfrak{A}.$$

It has the density

$$\varphi_k=\frac{d\nu_k}{d\mu}=\frac{\chi_k\varphi}{\nu(Y_k)},$$

where χ_k is the characteristic function of Y_k. Evidently, the sequence ν_k converges to ν in the fine topology, and (35) implies that $\rho(\nu_k,\mu)\to-\infty$. In addition, the condition $\mu(Y_k)<+\infty$ implies that $\rho(\nu_k,\mu)>-\infty$.

Fix an arbitrary $\varepsilon > 0$ and an arbitrary fine neighborhood $O(\nu)$. Choose k so large that $\nu_k \in O(\nu)$ and simultaneously $\rho(\nu_k, \mu) < -1/\varepsilon$. In the case of finite Kullback action, estimate (14) is already proven. Apply it to the pair of measures ν_k and μ, the neighborhood $O(\nu)$ of the measure ν_k, and the number $\varepsilon' = -1/\varepsilon - \rho(\nu_k, \mu)$:

$$\mu^n\{x \in X^n \mid \delta_{x,n} \in O(\nu)\} \geq e^{-n(\rho(\nu_k,\mu)+\varepsilon')} = e^{n/\varepsilon},$$

provided n is large enough. This is exactly estimate (14) for the case $\rho(\nu, \mu) = -\infty$. □

7. Proof of Theorems 4 and 5

Proof of Theorem 4. Under conditions of Theorem 4, the Kullback action is well-defined. Using the definition of well-definiteness and Theorem 2, we can choose a subset $X_\varepsilon \in \mathfrak{A}$ with $\mu(X_\varepsilon) < +\infty$, such that, for any set $Y \in \mathfrak{A}$ that contains X_ε and has a finite measure $\mu(Y)$, one of the following holds:

(a) $|\rho(\nu_Y, \mu) - \rho(\nu, \mu)| < \varepsilon$ in the case of finite $\rho(\nu, \mu)$,
(b) $\rho(\nu_Y, \mu) > 1/\varepsilon$ in the case $\rho(\nu, \mu) = +\infty$,
(c) $\rho(\nu_Y, \mu) < -1/\varepsilon$ in the case $\rho(\nu, \mu) = -\infty$.

Now, estimates (16) and (17) follow from the corresponding estimates of Theorem 1.

In addition, estimate (18) comes from estimate (7) of Theorem 1. To see this, it is enough to choose a set $Y \in \mathfrak{A}$ with $\mu(Y) < +\infty$ such that, along with one of the conditions (a)–(c), it satisfies the condition $\nu_Y \in O(\nu)$. □

Proof of Theorem 5. Suppose that the measure ν is absolutely continuous with respect to μ and has a density $\varphi = d\nu/d\mu$, but the Kullback action $\rho(\nu, \mu)$ is ill-defined. Consider the set

$$X_0 = \{x \in X \mid \varphi(x) \geq 1\}.$$

Obviously, $\mu(X_0) \leq \nu(X_0) \leq 1$. In addition, since the Kullback action is ill-defined,

$$\int_{X_0} \varphi \ln \varphi \, d\mu = +\infty.$$

For any measurable set $Y \supset X_0$ with $\mu(Y) < +\infty$, the corresponding conditional distribution ν_Y has the density $\chi_Y \varphi / \nu(Y)$. Therefore,

$$\rho(\nu_Y, \mu) = \int_Y \frac{\varphi}{\nu(Y)} \ln \frac{\varphi}{\nu(Y)} \, d\mu = +\infty,$$

and hence estimate (19) follows from estimate (6) of Theorem 1.

Now consider the case when the measure ν is not absolutely continuous with respect to μ. Then, there exists $X_0 \in \mathfrak{A}$ with $\mu(X_0) = 0$ and $\nu(X_0) > 0$. Suppose that a set $Y \in \mathfrak{A}$ with $\mu(Y) < +\infty$ contains X_0. Obviously, the conditional distribution ν_Y is not absolutely continuous with respect to μ and hence $\rho(\nu_Y, \mu) = +\infty$. Thus, we can apply Theorem 1 to the measures ν_Y, μ on the space Y and obtain (19). □

8. The Case of Finitely Additive Probability Distributions ν

The necessity of consideration of finitely additive probability distributions ν is caused by the fact that they may happen to be accumulation points for some sequences of empirical measures. Thus, to make the description of the empirical measures distribution complete, we should obtain the estimates similar to (13) and (14) for finitely additive probability distributions ν as well.

In fact, this can be done, and the principal result is that Theorems 3 and 5 *still hold true for finitely additive probability distributions* ν*, provided the Kullback action* $\rho(\nu, \mu)$ *is defined by* (10). In addition, in that case, $\rho(\nu, \mu)$ may take values $+\infty$ or $-\infty$ only, and the both are possible.

The transition from countably additive distributions to only finitely additive ones is not trivial. First of all, we should adapt some previous definitions to the new setting.

Denote by $N_1(X)$ the set of all finitely additive probability measures on $(X, \mathfrak{A})$. Each $\nu \in N_1(X)$ is naturally identified with a positive normalized linear functional on the space of bounded measurable functions $B(X)$ (i.e., a functional that takes nonnegative values on nonnegative functions and the unit value on the unit function). Using this identification, we denote the integral of $f \in B(X)$ with respect to $\nu \in N_1(X)$ as $\nu[f]$. In addition, for bounded above functions $f \in \bar{B}(X)$, let us define $\nu[f]$ as

$$\nu[f] = \lim_{c \to -\infty} \nu[f \vee c], \qquad f \vee c = \max\{f, c\}.$$

Thus, for $f \in \bar{B}(X)$, the value $\nu[f]$ belongs to the interval $[-\infty, +\infty)$. Similarly, for a measurable function f that is bounded from below, put

$$\nu[f] = \lim_{c \to +\infty} \nu[f \wedge c], \qquad f \wedge c = \min\{f, c\}.$$

Now, we define the Kullback action $\rho(\nu, \mu)$ for the case when $\nu \in N_1(X)$ and $\mu \in M_\sigma(X)$:

$$\rho(\nu, \mu) = \begin{cases} +\infty, & \text{if } \exists\, A \in \mathfrak{A}\colon\ \mu(A) = 0,\ \nu(A) > 0, \\ \sup\limits_{\psi \in \bar{B}(X)} \{\nu[\psi] - \lambda(\psi, \mu)\}, & \text{if } \nexists\, A \in \mathfrak{A}\colon\ \mu(A) = 0,\ \nu(A) > 0. \end{cases} \tag{36}$$

Obviously, this definition just duplicates (10).

Theorem 6. *If $\nu \in N_1(X)$ has no density with respect to $\mu \in M_\sigma(X)$, then $\rho(\nu, \mu)$ turns into $+\infty$ or $-\infty$. In particular, if μ is finite or ν is countably additive, then $\rho(\nu, \mu) = +\infty$.*

Let us introduce a fine topology on $N_1(X)$ by means of neighborhoods of the form

$$O(\nu) = \left\{\delta \in N_1(X) : \left|\delta[f_i] - \nu[f_i]\right| < \varepsilon,\ \ i = 1, \dots, k\right\}, \quad \nu \in N_1(X), \tag{37}$$

where $\varepsilon > 0$ and the functions $f_1, \dots, f_k \in \bar{B}(X)$ are such that all $\nu[f_i]$ are finite. Clearly, this definition is analogous to (5). Note that the bounded above functions in (37) may be replaced by bounded below or even nonnegative ones. This will not change the collection of neighborhoods (37).

Now, we reformulate Theorems 3 and 5 for the case of finitely additive distributions ν (note that Theorem 4 cannot be reformulated since $\rho(\nu, \mu)$ is well-defined, and hence ν is countably additive in it).

Theorem 7. *For any measures $\nu \in N_1(X)$, $\mu \in M_\sigma(X)$, and number $\varepsilon > 0$, there exists a fine neighborhood $O(\nu) \subset N_1(X)$ such that*

$$\mu^n\left\{x = (x_1, \dots, x_n) \in X^n \;\middle|\; \delta_{x,n} \in O(\nu)\right\} \le e^{-n(\rho(\nu,\mu)-\varepsilon)}, \qquad n = 1, 2, 3, \dots \tag{38}$$

On the other hand, for any measures $\nu \in N_1(X)$, $\mu \in M_\sigma(X)$, number $\varepsilon > 0$, and any fine neighborhood $O(\nu) \subset N_1(X)$, the following estimate holds for all large enough n:

$$\mu^n\left\{x = (x_1, \dots, x_n) \in X^n \;\middle|\; \delta_{x,n} \in O(\nu)\right\} \ge e^{-n(\rho(\nu,\mu)+\varepsilon)}. \tag{39}$$

If $\rho(\nu, \mu) = +\infty$, then the difference $\rho(\nu, \mu) - \varepsilon$ in (38) should be replaced by $1/\varepsilon$, and if $\rho(\nu, \mu) = -\infty$, then the sum $\rho(\nu, \mu) + \varepsilon$ in (39) should be replaced by $-1/\varepsilon$.

A measure $\nu \in N_1(X)$ will be called *proper* with respect to a measure $\mu \in M_\sigma(X)$ if, for any $\varepsilon > 0$, there exists a set $A \in \mathfrak{A}$ such that $\mu(A) < +\infty$ and $\nu(A) > 1 - \varepsilon$. If, on the contrary, there exists an $\varepsilon > 0$ such that the inequality $\nu(A) > 1 - \varepsilon$ implies $\mu(A) = +\infty$, then the measure ν will be called *improper*

with respect to μ. Obviously, in the case of finite μ, all measures $\nu \in N_1(X)$ are proper, and, in the case of σ-finite μ, all countably additive measures $\nu \in N_1(X)$ are proper.

Theorem 8. *Suppose that for some measures $\nu \in N_1(X)$ and $\mu \in M_\sigma(X)$, the Kullback action $\rho(\nu, \mu)$ is ill-defined, and the measure ν is proper with respect to μ. Then, there exists a set $X_0 \in \mathfrak{A}$ with $\mu(X_0) < +\infty$, such that, for any $Y \in \mathfrak{A}$ containing X_0 and having a finite measure $\mu(Y)$, and any $\varepsilon > 0$, there exists a weak neighborhood $O(\nu_Y) \subset N_1(Y)$ satisfying the estimate*

$$\mu^n\{x = (x_1, \dots, x_n) \in Y^n \mid \delta_{x,n} \in O(\nu_Y)\} \le e^{-n/\varepsilon}, \qquad n \in \mathbb{N}. \tag{40}$$

9. Proof of Theorem 6

Lemma 9. *Suppose that a measure $\nu \in N_1(X)$ is proper with respect to $\mu \in M_\sigma(X)$, and, for any $\varepsilon > 0$, there exists $\delta > 0$, such that $\mu(A) < \delta$ implies $\nu(A) < \varepsilon$. Then, ν is countably additive and absolutely continuous with respect to μ.*

Proof. Construct a sequence of embedded measurable sets $A_1 \subset A_2 \subset A_3 \subset \cdots$ such that all of them have finite measures $\mu(A_n)$, satisfy the condition $\nu(A_n) > 1 - 1/n$, and their union is the whole X.

The restriction of μ to each A_n is finite and continuous: if a sequence of embedded measurable sets $A_n \supset B_1 \supset B_2 \supset B_3 \supset \cdots$ has an empty intersection, then $\mu(B_k) \to 0$. The assumption of Lemma 9 implies that the restriction of ν to A_n is continuous as well. It is known that the continuity of a finite measure is equivalent to its countable additivity. Then, the restriction of ν to each A_n is countably additive. Since ν is proper, we have $\nu(B) = \lim_{n\to\infty} \nu(B \cap A_n)$ for any measurable B. It follows that ν is countably additive on the whole X (this may be proven in the same way as the countable additivity of a σ-finite measure). □

Proof of Theorem 6. It follows from (36) that either $\rho(\nu, \mu) = +\infty$ or

$$\rho(\nu, \mu) = \sup_{\psi \in \bar{B}(X)} \{\nu[\psi] - \lambda(\psi, \mu)\}. \tag{41}$$

In the first case, the assertion of Theorem 6 is valid. Therefore, it is enough to consider the case when the Kullback action is defined by formula (41).

By the assumption of Theorem 6, the measure $\nu \in N_1(X)$ has no density with respect to μ. Then, Lemma 9 guarantees validity of at least one of the following two conditions:

(a) there exists a positive ε, such that, for any $\delta > 0$, one can choose a measurable set A_δ such that $\mu(A_\delta) < \delta$ and $\nu(A_\delta) \ge \varepsilon$;
(b) the measure ν is improper with respect to μ.

Suppose that (a) is valid. If $\rho(\nu, \mu) > -\infty$, then (41) implies existence of a function $\psi \in \bar{B}(X)$ such that $\nu[\psi] - \lambda(\psi, \mu) > -\infty$. Fix a number $t > 0$ and consider the family of functions $\psi_\delta = \psi + t\chi_\delta$, where χ_δ is the characteristic function of the set A_δ. When $\delta \to 0$, we have

$$\nu[\psi_\delta] - \lambda(\psi_\delta, \mu) \ge \nu[\psi] + t\varepsilon - \ln\left\{\int_X e^{\psi}\, d\mu + \int_{A_\delta} e^{\psi + t}\, d\mu\right\} \longrightarrow \nu[\psi] + t\varepsilon - \lambda(\psi, \mu). \tag{42}$$

From the arbitrariness of t, (41) and (42), it follows that $\rho(\nu, \mu) = +\infty$.

Now assume that (b) is valid. Consider any function $\psi \in \bar{B}(X)$ such that $\nu[\psi] > -\infty$. Define the sets $A_n = \{x \in X \mid \psi(x) \ge -n\}$. The condition $\nu[\psi] > -\infty$ implies $\nu(A_n) \to 1$. Since the measure ν is improper, it follows that $\mu(A_n) = +\infty$ for all large enough n. Then,

$$\lambda(\psi, \mu) = \ln \int_X e^{\psi}\, d\mu \ge \ln \int_{A_n} e^{-n}\, d\mu = -n + \ln \mu(A_n) = +\infty,$$

and hence

$$\rho(\nu,\mu) = \sup_{\psi\in\bar{B}(X)} \{\nu[\psi] - \lambda(\psi,\mu)\} = -\infty.$$

Recall that if μ is finite, then ν is proper, and hence alternative (b) cannot take place. In addition, for finite μ one has $\rho(\mu,\mu) > -\infty$. Thus, (a) implies $\rho(\nu,\mu) = +\infty$. If ν is countably additive and has no density with respect to μ, then the first case of (36) takes place, according to which $\rho(\nu,\mu) = +\infty$ as well. □

10. Proof of Theorems 7 and 8

The proof for the first part of Theorem 7 is exactly the same as for the first part of Theorem 3, so we omit it. If $\nu \in M_1(X)$, then the second part of Theorem 7 follows from the second part of Theorem 3. Thus, it remains to consider the case $\nu \in N_1(X) \setminus M_1(X)$.

Let $\mathfrak{B}$ be some σ-field of subsets of X. We will call it *discrete* if it is generated by a countable or finite partition of X.

Lemma 10. *For any measure $\nu \in N_1(X)$ and any its fine neighborhood $O(\nu)$, there exists a discrete σ-subfield $\mathfrak{B} \subset \mathfrak{A}$ such that*

(*a*) *the restriction of ν to $\mathfrak{B}$ is countably additive;*
(*b*) *there exists a fine neighborhood $O'(\nu) \subset O(\nu)$ generated by $\mathfrak{B}$-measurable functions;*
(*c*) *if the measure ν is proper with respect to $\mu \in M_\sigma(X)$, then the σ-field $\mathfrak{B}$ mentioned above can be chosen in such a way that each of its atoms has a finite measure μ.*

Proof. A base for the fine topology on $N_1(X)$ is formed by the neighborhoods

$$O(\nu) = \big\{\delta \in N_1(X) : \big|\delta[f_i] - \nu[f_i]\big| < 3\varepsilon,\ \ i = 1,\dots,m\big\}, \qquad \varepsilon > 0,$$

where f_i are measurable nonnegative functions on $(X,\mathfrak{A})$ with $\nu[f_i] < +\infty$. Let us prove the Lemma for a neighborhood of this sort.

Define the step-functions $g_i = \varepsilon[f_i/\varepsilon]$, where $[\,\cdot\,]$ denotes the integer part of a number, and the neighborhood

$$O'(\nu) = \big\{\delta \in N_1(X) : \big|\delta[g_i] - \nu[g_i]\big| < \varepsilon,\ \ i = 1,\dots,m\big\}.$$

Evidently, $|g_i - f_i| \le \varepsilon$, and, for each $\delta \in O'(\nu)$, we have

$$\big|\delta[f_i] - \nu[f_i]\big| \le \big|\delta[f_i] - \delta[g_i]\big| + \big|\delta[g_i] - \nu[g_i]\big| + \big|\nu[g_i] - \nu[f_i]\big| < 3\varepsilon.$$

It follows that $O'(\nu) \subset O(\nu)$.

To each integer vector $k = (k_1,\dots,k_m) \in \mathbb{Z}^m$, assign the set

$$X_k = \big\{x \in X \,\big|\, g_i(x) = k_i\varepsilon,\ \ i = 1,\dots,m\big\}.$$

These sets form a countable measurable partition of X and generate the desired discrete σ-subfield $\mathfrak{B}$. The functions g_i are $\mathfrak{B}$-measurable.

Note that, for any $C > 0$, we have

$$\nu\{x \in X \mid g_i(x) \ge C\} \le \frac{\nu[g_i]}{C}.$$

Thus, when C goes to $+\infty$,

$$\sum_{\substack{k_i\le C,\\ i\le m}} \nu(X_k) \ge 1 - \sum_{i=1}^{m} \nu\{x \in X \mid g_i(x) \ge C\varepsilon\} \ge 1 - \sum_{i=1}^{m} \frac{\nu[g_i]}{C\varepsilon} \ \to\ 1.$$

It follows that the restriction of ν to the σ-field $\mathfrak{B}$ is countably additive.

Assume that the measure ν is proper with respect to μ. In this case, we can construct a countable partition of X into subsets $Y_l \in \mathfrak{A}$ such that $\mu(Y_l) < +\infty$ and $\nu(Y_1 \sqcup \cdots \sqcup Y_l) \to 1$ as $l \to +\infty$. The latter condition implies the equality $\nu(X_k) = \sum_l \nu(X_k \cap Y_l)$. Therefore, the restriction of ν to the σ-field generated by the atoms $X_k \cap Y_l$ is countably additive. This σ-field may be treated as $\mathfrak{B}$. By construction, its atoms have finite measure μ. □

Let us finish the proof of Theorem 7. It remains to obtain estimate (39) for $\nu \in N_1(X) \setminus M_1(X)$. In this situation, the measure ν has no density with respect to μ, and, according to Theorem 6, we have the alternative: either $\rho(\nu, \mu) = +\infty$ or $\rho(\nu, \mu) = -\infty$. In the first case, estimate (39) is trivial. Thus, it is enough to consider the second case $\rho(\nu, \mu) = -\infty$.

Suppose the measure ν is proper with respect to μ and $\rho(\nu, \mu) = -\infty$. We can apply Lemma 10 to ν and construct the corresponding discrete σ-subfield $\mathfrak{B} \subset \mathfrak{A}$ and fine neighborhood $O'(\nu) \subset O(\nu)$. Denote by $\bar{\nu}$ and $\bar{\mu}$ the restrictions of ν and μ to $\mathfrak{B}$. By Lemma 10, they are countably additive. From definition (36), it follows that if $\bar{\mu}(A) = 0$ for some $A \in \mathfrak{B}$, then $\bar{\nu}(A) = 0$ as well (since otherwise $\rho(\nu, \mu) = +\infty$). Thus, the distribution $\bar{\nu}$ on $\mathfrak{B}$ is absolutely continuous with respect to $\bar{\mu}$.

Recall that by definition,

$$\rho(\nu, \mu) = \sup_{\psi \in \bar{B}(X)} \{\nu[\psi] - \lambda(\psi, \mu)\} = -\infty,$$

where $\bar{B}(X)$ is the set of all bounded above $\mathfrak{A}$-measurable functions. The same is true for all bounded above $\mathfrak{B}$-measurable functions, and hence $\rho(\bar{\nu}, \bar{\mu}) = -\infty$ as well. Since $\bar{\nu}$ is absolutely continuous with respect to $\bar{\mu}$, the second part of Theorem 7 for $\bar{\nu}$ and $\bar{\mu}$ is proven. It implies the estimate

$$\bar{\mu}^n \{x \in X^n \mid \delta_{x,n} \in O'(\nu)\} \geq e^{n/\varepsilon}$$

for all large enough n. Due to the inclusion $O'(\nu) \subset O(\nu)$, we obtain (39).

Consider the case of improper ν. We can apply Lemma 10 and construct the corresponding discrete σ-subfield $\mathfrak{B}$ and a fine neighborhood $O'(\nu) \subset O(\nu)$ generated by $\mathfrak{B}$-measurable functions. The field $\mathfrak{B}$ is generated by a certain denumerable partition $X = X_1 \sqcup X_2 \sqcup X_3 \sqcup \cdots$. Change numeration of the sets X_i so that $\nu(X_1) \geq \nu(X_2) \geq \nu(X_3) \geq \cdots$. Put $Y_k = X_1 \sqcup X_2 \sqcup \cdots \sqcup X_k$ and denote by ν_k the conditional distribution of ν on Y_k. Due to the countable additivity, $\nu(Y_k) \to 1$ and $\nu_k \in O'(\nu)$ for all large enough k. In addition, the improperness of ν implies that $\mu(Y_k) = +\infty$ for all large enough k.

Fix such a large k that $\nu_k \in O'(\nu)$, and, at the same time, $\mu(Y_k) = +\infty$. The latter implies $\mu(X_i) = +\infty$ for at least one $i \leq k$. Without loss of generality, we may assume that $\nu_k(X_i) > 0$ for all $i \leq k$. Obviously, for any large enough n, there exists a sequence $y = (y_1, \ldots, y_n) \in Y_k^n$ such that the empirical measure $\delta_{y,n}$ is so close to ν_k that $\delta_{y,n} \in O'(\nu)$ and each of the sets $X_1, \ldots, X_k$ contains at least one of the points $y_1, \ldots, y_n$. Define positive integers i_j in such a way that $y_j \in X_{i_j}$ for $j = 1, \ldots, n$. Then, $\mu(X_{i_j}) > 0$ for all $j = 1, \ldots, n$ (since otherwise $\rho(\nu, \mu) = +\infty$) and $\mu(X_{i_j}) = +\infty$ for at least one j. Therefore,

$$\mu^n \{x \in Y_k^n \mid \delta_{x,n} \in O'(\nu)\} \geq \mu^n \{x \in Y_k^n \mid x_j \in X_{i_j},\ j = 1, \ldots, n\} = \prod_{j=1}^{n} \mu(X_{i_j}) = +\infty,$$

and thereby estimate (39) is completely proven. □

Proof of Theorem 8. If $\nu \in M_1(X)$, then the assertion of Theorem 8 follows from Theorem 5.

Let $\nu \in N_1(X) \setminus M_1(X)$. Then, ν is not absolutely continuous with respect to μ.

Since ν is proper, by Lemma 9, there exists an $\varepsilon_0 > 0$ such that, for any positive integer n, there exists $A_n \in \mathfrak{A}$ satisfying $\mu(A_n) < 2^{-n}$ and $\nu(A_n) \geq \varepsilon_0$. Set $X_0 = \bigcup_n A_n$.

Suppose a set $Y \in \mathfrak{A}$ with $\mu(Y) < +\infty$ contains X_0. Then, the conditional distribution ν_Y is not absolutely continuous with respect to μ. On the other hand, (36) and the conditions $\mu(Y) < +\infty$ and

$\nu_Y(X \setminus Y) = 0$ imply the inequality $\rho(\nu_Y, \mu) > -\infty$. Hence, $\rho(\nu_Y, \mu) = +\infty$ by Theorem 6. In this case, estimate (40) follows from estimate (6) of Theorem 1. □

Author Contributions: A general idea of the research was suggested by Victor Bakhtin. Theorems 2, 3, 4, and 5 were obtained in collaboration of the authors. Theorems 6, 7, and 8 and their proofs belong to Victor Bakhtin. Both authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Billingsley, P. *Ergodic Theory and Information*; Wiley: New York, NY, USA, 1965; 204p.
2. Algoet, P.H.; Cover, T.M. A sandwich proof of the Shannon—McMillan—Breiman theorem. *Ann. Probab.* **1988**, *16*, 899–909.
3. Sanov, I.N. On the probability of large deviations of random variables. *Matematicheskii Sbornik* **1957**, *42*, 11–44. (In Russian)
4. Groeneboom, P.; Oosterhoff, J.; Ruymgaart, F.H. Large deviation theorems for empirical probability measures. *Ann. Probab.* **1979**, *7*, 553–586.
5. Jain, N. An introduction to large deviations. *Lect. Notes Math.* **1985**, *1153*, 273–296.
6. Deuschel, J.-D.; Stroockn, D.W. *Large Deviations: Pure and Applied Mathematics*; Academic Press: Boston, MA, USA, 1989; Volume 137.
7. Borovkov, A.A.; Mogul'skii, A.A. On large deviation principles in metric spaces. *Sib. Math. J.* **2010**, *51*, 989–1003.
8. Acosta, A. On large deviations of empirical measures in the τ-topology. *Stud. Appl. Probab.* **1994**, *31A*, 41–47.
9. Bakhtin, V.I. Spectral potential, Kullback action, and large deviations of empirical measures on measurable spaces. *Theory Probab. Appl.* **2015**, *59*, 535–544.
10. Bakhtin, V.I. Spectral potential, Kullback action, and large deviation principle for finitely-additive measures. *Tr. Inst. Mat.* **2015**, *23*, 11–23. (In Russian)
11. Donsker, M.D.; Varadhan, S.R.S. Asymptotic evaluation of certain Markov process expectations for large time, I. *Commun. Pure Appl. Math.* **1975**, *28*, 1–47.
12. Donsker, M.D.; Varadhan, S.R.S. Asymptotic evaluation of certain Markov process expectations for large time, II. *Commun. Pure Appl. Math.* **1975**, *28*, 279–301.
13. Donsker, M.D.; Varadhan, S.R.S. Asymptotic evaluation of certain Markov process expectations for large time, III. *Commun. Pure Appl. Math.* **1976**, *29*, 389–461.
14. Donsker, M.D.; Varadhan, S.R.S. Asymptotic evaluation of certain Markov process expectations for large time, IV. *Commun. Pure Appl. Math.* **1983**, *36*, 183–212.
15. Varadhan, S.R.S. Large deviations. *Ann. Probab.* **2008**, *36*, 397–419.
16. Freidlin, M.I.; Wentzell, A.D. *Random Perturbations of Dynamical Systems*, 3rd ed.; Springer: Berlin, Germany, 2012; 458p.
17. Borovkov, A.A.; Mogul'skii, A.A. Large deviation principles for random walk trajectories. I. *Theory Probab. Appl.* **2012**, *56*, 538–561.
18. Borovkov, A.A.; Mogul'skii, A.A. Large deviation principles for random walk trajectories. II. *Theory Probab. Appl.* **2013**, *57*, 1–27.
19. Borovkov, A.A.; Mogul'skii, A.A. Large deviation principles for random walk trajectories. III. *Theory Probab. Appl.* **2014**, *58*, 25–37.
20. Borovkov, A.A. *Mathematical Statistics*; Gordon & Breach: Amsterdam, The Netherlands, 1998; 570p.
21. Sokol, E.E. A generalization of McMillan's theorem on the case of countable alphabet. *Tr. Inst. Mat.* **2015**, *23*, 115–122. (In Russian)
22. Sokal, E.E. On the informational sense of entropy in the case of countable alphabet. *Vestn. Beloruss. Gos. Univ. Ser. 1 Fiz. Mat. Inform.* **2016**, *1*, 96–100. (In Russian)